Communicating with a Semi-Autonomous Robot Combining Natural Language and Gesture

Dennis Perzanowski, William Adams, and Alan C. Schultz

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Washington, DC 20375-5337
<dennisp | adams | schultz@aic.nrl.navy.mil>
<http:// www.aic.nrl.navy.mil/ ~dennisp | ~adams | ~schultz>
voice phone: 202-767-9005
fax: 202-767-3172

Abstract

The Intelligent Multimodal Multimedia and the Adaptive Systems Groups at the Navy Center for Applied Research in Artificial Intelligence have been investigating a natural language and gesture interface to a mobile robot. Our interface utilizes robust natural language understanding and resolves some of the ambiguities in natural language by means of gesture input. The natural language and gestural information is integrated with knowledge of a particular environment and appropriate robotic responses are produced.

So-called “deictic” elements or objects (e.g. “this chair,” “that table,” “him” or “her”) and directional elements (e.g. “over there,” “my left” and “your right”) when parsed by a natural language system can be comprehensible, but mean nothing if the utterance is unaccompanied by gesture. A command such as "Go/Move over there” is ambiguous without an appropriate gesture to indicate some place in the environment to which to move. Moreover, a command such as “Turn left fifteen degrees” can be confusing if an inappropriate or contradictory gesture is perceived. This interface handles both natural language ambiguity and appropriate or inappropriate (contradictory) gestures.

Introduction

Our research implementing a natural language and gestural interface to a semi-autonomous robot is based on two assumptions. The first, or linguistic, assumption is that certain types of ambiguity in natural language can be resolved when gestures are incorporated in the input. For example, a sentence such as “Go over there” is devoid of meaning unless it is accompanied by a gesture indicating the place where the speaker wishes the hearer to move. Furthermore, while gestures are an integral part of communication [1], our second, or gestural, assumption is that stylized or symbolic gestures place a heavier burden on the human, frequently requiring a learning period, since such gestures tend to be arbitrary in nature. Natural gestures, i.e. gestures that do not require learning and which any human might produce as a natural co-occurrence to a particular verbal command, are simpler means of imparting certain kinds of information in human-computer interaction. With systems that have fairly robust vision capabilities, natural gestures obviate the need for additional interactive devices, such as computer terminals, touchscreens, or data gloves. So from a linguistic and gestural standpoint, certain utterances, such as those that involve movement or location information, can be disambiguated by means of natural, accompanying gesture [2].

For this study, we limit ourselves to two types of commands: commands that involve direction, e.g. “Turn left,” and those that involve locomotion, e.g. “Go over there.” For such commands, environmental conditions permitting, people communicate with each other by pointing to objects in their surroundings, or gesturing in the specified direction. Granted, if the environment or meteorological conditions are not favorable, as for example when it is too dark to see or if foggy or heavy precipitation prevails, humans may rely on other methods to communicate, which will not concern us here. However, given a more or less ideal environment, human to human communication typically involves the use of natural language and gesture, and it is this type of interaction that we have emulated in our human-computer interface to a semi-autonomous robot.

For the kinds of interaction that we have outlined above, touchscreens or data gloves also allow humans to communicate and talk about deictic elements in various computer applications. So-called “deictic” elements are linguistic strings that refer to objects in the discourse which in turn usually refer to objects in the real world. For example, in the sentence “the box in the corner is red,” the subject of the sentence “the box in the corner” can be analyzed as a deictic element if one exists in the same environment as the speaker and/or hearer of this utterance. If the intended referent, namely “the box,” does not exist in this environment, either the speaker is playing some sort of linguistic trick, or the utterance is uninterpretable. More typically, deictic elements are characterized by the presence of such words as "this" or "that" for objects, as in the expressions, "this box is red," "that is a blue box," or "here" or "there" for locations, as in "bring it over here," or "the waypoint over there." We limit ourselves in this research to statements in which deictic elements exist in the world of the speaker and hearer.

Based on our initial assumptions, we have chosen to limit our consideration to so-called "natural" gestures. Granted, someone can disambiguate "Go/Move over there" by moving one's head in a particular direction, or by moving one's eyes. However, our system cannot currently handle such complex and minute movements for the purposes of disambiguation, and we limit ourselves to grosser movements, such as hand and arm movements.

The Natural Language and Gesture Interface

This work is based on research already underway at the Navy Center for Applied Research in Artificial Intelligence (NCARAI) [3]. To process verbal or audial commands issued to a semi-autonomous robot, the natural language interface (Figure 1) utilizes a 70-word speech vocabulary with an input range of approximately 11,000 utterances.

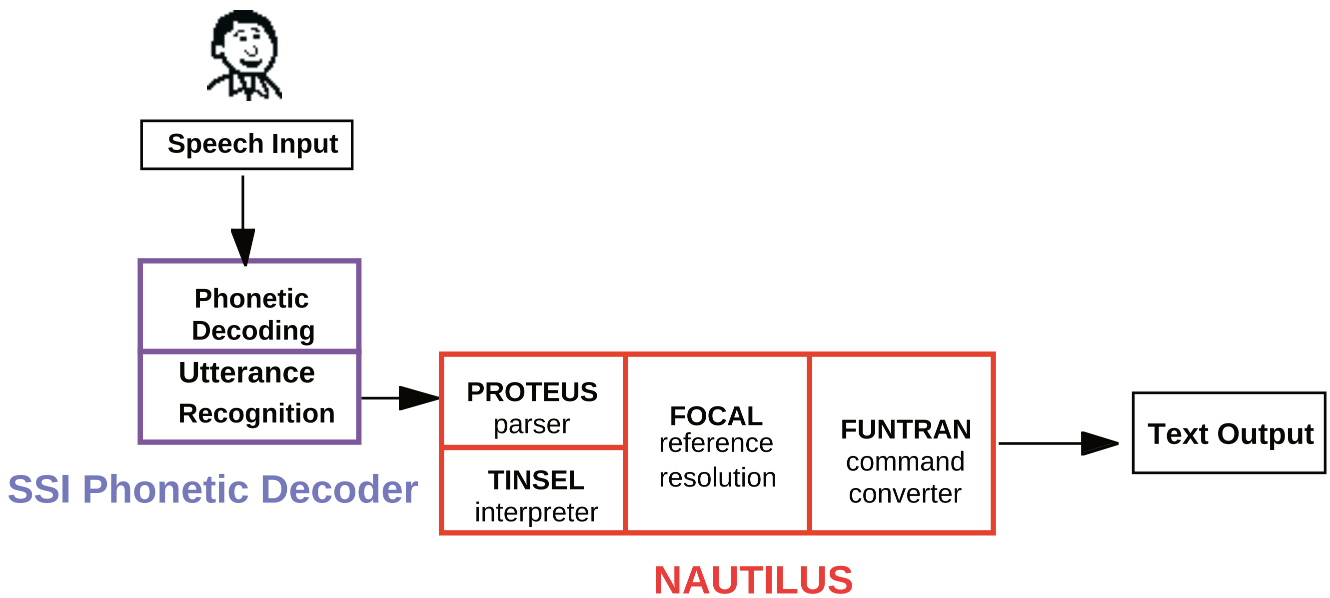

Figure 1: Natural language processing of command utterances

Natural Language Processing

The auditory signal is converted to a text string by the speech recognition system, a PE 200, manufactured by Speech Systems, Inc. The textual string is then parsed by our natural language processing system, NAUTILUS, developed in-house at NCARAI [4].

We have opted for robust natural language parsing, starting with syntactic analysis of the input string, since one of our project's research objectives [5] has been to port the grammar and parsing mechanism to various applications and domains.

During the parsing process, the syntactic string is semantically interpreted in several of the NAUTILUS modules, and the resulting semantic representation is submitted to the command module on the semi-autonomous robot for further processing (Figure 2).

The mapping of the semantic interpretation, the Lisp-like structure on the left-hand side of Figure 2, and the perceived gesture, the numerical string on the right-hand side of Figure 2, to produce some kind of action is best understood after a consideration of the gesture processing, to which we now turn.

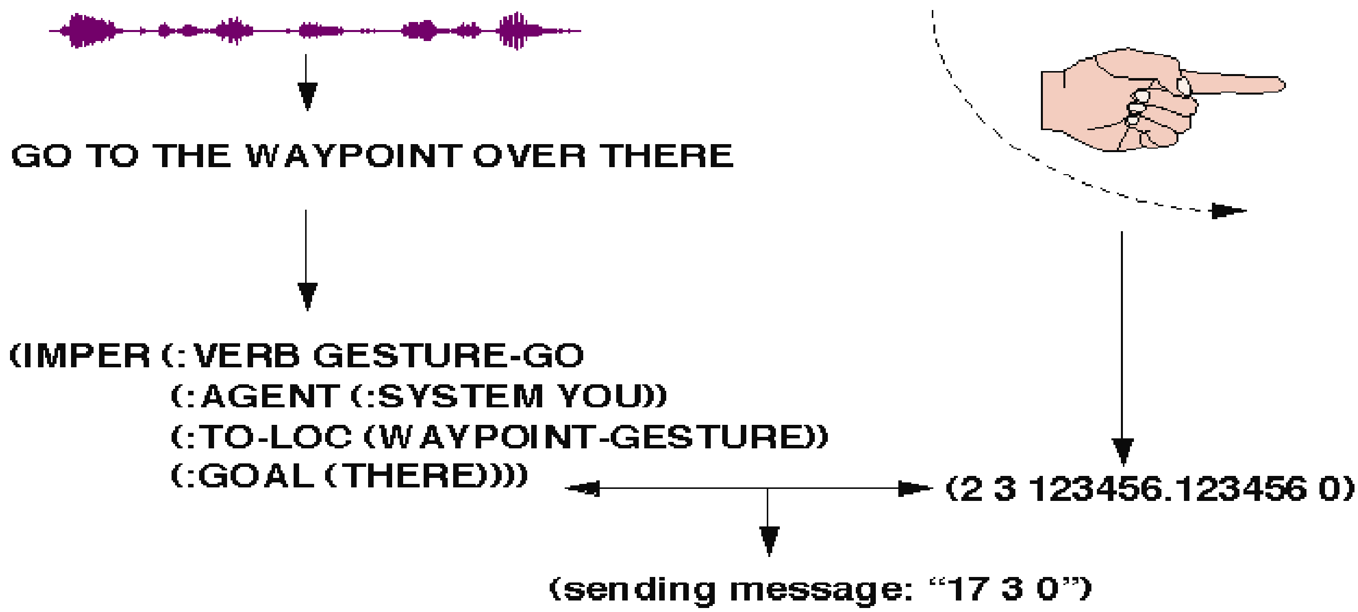

Figure 2: Integration of natural language processing and gesture analysis

Gesture Processing

The semi-autonomous robot in this investigation is a mobile Nomad 200, manufactured by Nomadic Technologies, Inc., and equipped with 16 Polaroid sonars and 16 active infrared sensors. It is capable of detecting and incorporating gesture into various verbal commands (Figure 3).

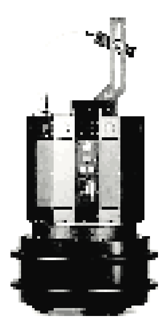

Figure 3: A Nomad 200 mobile robot with mounted camera

The robot is capable of sensing vectors and measured line segments that the human might gesture during the various commands. The robot is linked as a UNIX workstation via a radio ethernet connection to the natural language processing modules and to the command modules of the robot.

A process running on the robot is used to determine the gestures given by the human user. The gestures are detected with a structured light rangefinder which emits a horizontal plane of laser light 30 inches above the floor. A camera mounted above the laser is fitted with a filter tuned to the laser frequency. The camera observes the intersection of the laserlight with any objects in the room, and the bright pixels in the camera's image are mapped to XY coordinates.

Periodically, the data points from one camera frame are used to compute an average distance from the objects seen. The points are then sorted into clusters, and any cluster sufficiently closer than the average and of appropriate size is designated as a hand. Hand locations are stored for multiple frames until no hands are found or a maximum number of frames are used.

The hand locations across the frames are ordered into one or two trajectories. Trajectories are built incrementally by grouping each hand location of each frame with the trajectory with which it best aligns.

Completed trajectories are checked to see if they are in motion or are stationary, and then logically compared to determine if the overall gesture is valid and if so which gesture was made. The valid gestures are queued and when the multimodal software needs to check for a gesture, it queries the gesture process which returns the most recent gesture from the queue. This is the string, a kind of data structure, on the right-hand side of Figure 2.

Mapping Speech Input with a Gesture

Returning to the natural language side of processing the input, an utterance, such as "Go/Move over there", is inherently ambiguous without additional information. Cues must be supplied, namely an accompanying gesture for the sentence to be completely understood. The natural language system can parse the utterance and give an adequate semantic interpretation to it, but without a visual cue, such as a hand gesture in a particular direction, the sentence is ultimately meaningless. Given the vision capability on the Nomad 200 robot and the processing as outlined above, when the sensors on the robot detect a vector within the limitations of its light striping sensor, and a command is sent to move in some direction, a query is made of the gesture process on the robot to see if some gesture has been perceived. The two are then mapped to a message, the final output in Figure 2 which is then sent back to the robot in order to produce an appropriate action. The mapping of the speech input and the perceived gesture is a function of the appropriateness or inappropriateness and the presence or absence of a gesture during the speech input.

Additional Types of Input

To handle specific objects from the environment in the various commands which the robot is capable of processing, we have introduced these objects into the semantic component of the natural language processing system. Thus, when a sentence such as “Go to waypoint two” is uttered, and whether or not an object is pointed to in the environment, a meaningful utterance is obtained. The knowledge base of the robot is consulted, and given the fact that an appropriate object (waypoint two) exists in the semantics of the natural language component and in the robot’s knowledge base, the robot then moves toward the known object in the room. In this case, since a referent in the real world exists and is known, and has been uniquely identified in the speech signal, a gesture is redundant, and can be ignored. If the human utters the sentence “Go to the waypoint over there,” after natural language processing successfully parses and interprets the utterance, it queries the robot’s knowledge base to check if there is an object of such a description as a “waypoint” located in that area of the room, assuming a gesture has been made. If the query receives an affirmative response, the robot moves off to the intended goal. If the human points to some location in the room where there is no known waypoint, then an appropriate error response is produced, such as “I don’t understand. There is no waypoint in that direction.” Of course, if no gesture was perceived, the robot responds appropriately that one is required.

Additional commands, such as “Back up/Move forward this far” are handled in a similar fashion. If the camera mounted on the top of the robot senses a measured line segment, the robot moves the gestured distance in the intended direction; however, without an appropriate gesture, such a sentence evokes an error response, such as “How far?”.

Directional commands are treated in much the same way. For example, the robot can be told to turn in any direction an arbitrary number of degrees.[1] Such utterances as “Turn 30 degrees to the left”, “Turn to your left/right” or “Turn to my left/right” produce appropriate robotic responses. However, if the human issues a contradictory gesture while uttering these commands, the robot responds accordingly, stating that a contradictory gesture was perceived, and no further action is taken at that time. Likewise, if the robot is told to “Turn this/that way,” a gesture must be perceived; otherwise, an error message results, stating that no gesture was observed and one is required with such a command.

Two Constraints of the Current System

As we continued to process directional commands involving turning and moving toward objects, we realized that some of the directional capabilities of our system were being severely constrained because of certain physical constraints imposed on us by our speech recognition system.

The system that we currently employ, the SSI PE 200, requires that the human user be tethered to the speech recognition device. It further requires the user to wear a headset that is directly connected by a cable to the speech recognition system. When any orientation of speaker and robot, other than face-to-face orientation, resulted, we were forced to maneuver the robot so that its vision system was capable of seeing the user and any gestures produced. At times, this required some strategic re-alignment of robot and human user. Also, a command such as “Turn to my left/right” suddenly takes on an entirely different meaning when speaker and hearer are no longer oriented face-to-face. With a mobile headset, the user will be able to move more freely in the environment, thereby increasing interaction.

Furthermore, because of the vision capabilities that are currently used by our robot, we were limited to vectoring gestures or gestures that segmented a line. A more sophisticated visual system, such as one that is capable of discriminating hand shapes and detecting skin tone [6,7], will provide greater opportunities for us to explore different kinds of hand gestures in conjunction with verbal commands.

Conclusions

Our goal has been to develop a natural language and gestural interface to a semi-autonomous robot. The use of natural language and gesture in the interface is based on our assumptions that while natural language is ambiguous, gestures disambiguate certain information in speech, and secondly, natural gestures are more easily used by humans giving directive and locomotive commands to a mobile robot. Our interface does not require the user to wear any

[1] As anyone involved with speech recognition systems knows, numbers are still extremely difficult to understand and pose major problems. However, since the focus of our work here has been on integrating natural language and gesture and not totally on speech understanding, we await a more sophisticated and advanced speech recognition system to continue our work.

special gear, nor to learn a series of symbolic gestures in order to interface with the robot, and while our corpus of commands is limited, it has been constrained to meet the limitations of the current vision system of the mobile robot. We believe this research exhibits how easily natural language and gesture can be used to interact with a semi-autonomous robot. In the future, we hope to expand the research by incorporating a more robust speech recognition system. This will enable us to increase the types of verbal information that can be input for commands to a mobile robot. We also wish to employ a more sophisticated vision system, such as one capable of detecting the human hand so that more complex gestures can be incorporated with the speech signal.

Acknowledgments

This work is funded in part by the Naval Research Laboratory and the Office of Naval Research.

References

[1] Kortenkamp, D; Huber, E.; and Bonasso, R.P. 1996. Recognizing and Interpreting Gestures on a Mobile Robot. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, 915-921. Menlo Park, CA: AAAI Press.

[2] Cassell, J. 1995. Speech, Action and Gestures as Context for On-going Task-oriented Talk. In Working Notes of the 1995 AAAI Fall Symposium on Embodied Language and Action, 20-25. Menlo Park, CA: AAAI Press.

[3] Yamauchi B.; Schultz, A.; Adams, W.; Graves, K.; Grefenstette, J.; and Perzanowski, D. 1997. ARIEL: Autonomous Robot for Integrated Exploration and Localization. In Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, 804-805. Menlo Park: CA: AAAI Press.

[4] Wauchope, K. 1994. Eucalyptus: Integrating Natural Language Input with a Graphical User Interface, NRL Technical Report, NRL/FR/5510--94-9711, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC.

[5] Wauchope K.; Everett, S.; Perzanowski, D.; and Marsh, E. 1997. Natural Language in Four Spatial Interfaces. In Proceedings of the Fifth Conference on Applied Natural Language Processing, 8-11. Washington, DC: Association for Computational Linguistics.

[6] Erenshteyn, R., P. Laskov, R. Foulds, L. Messing, and G. Stern. 1996. Recognition Approach to Gesture Language Understanding In Proceedings of the 13th International Conference on Pattern Recognition, 431-435. Los Alamitos, CA: IEEE Computer Society Press.

[7] Saxe, D. and R. Foulds. 1996. Toward Robust Skin Identification in Video Images. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 379-384. Los Alamitos, CA: IEEE Computer Society Press.